Where is the Value in High Frequency Trading? *

Álvaro Cartea† and José Penalva‡

February 28, 2011

Abstract

We analyze the impact of high frequency trading in financial markets based on a model with three types of traders: liquidity traders, market makers, and high frequency traders. Our four main findings are: i) The price impact of the liquidity trades is higher in the presence of the high frequency trader and is increasing with the size of the trade. In particular, we show that the high frequency trader reduces (increases) the prices that liquidity traders receive when selling (buying) their equity holdings. ii) Although market makers also lose revenue to the high frequency trader in every trade, they are compensated for these losses by a higher liquidity discount. iii) High frequency trading increases the volatility of prices. iv) The volume of trades doubles as the high frequency trader intermediates all trades between the liquidity traders and market makers. This additional volume is a consequence of trades which are carefully tailored for surplus extraction and are neither driven by fundamentals nor is it noise trading. In equilibrium, high frequency trading and traditional market making coexist as competition drives down the profits for new high frequency traders while the presence of high frequency traders does not drive out traditional market makers.

*We would like to thank Harrison Hong for his valuable comments and discussions. We are also grateful to Andrés Almazán, Gene Amromin, Michael Brennan, Pete Kyle and Eduardo Schwartz for their comments. We also thank seminar participants at CEMFI. The usual caveat applies. We welcome comments, including references we have inadvertently missed. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banco de España.

†alvaro.cartea@uc3m.es, Universidad Carlos III de Madrid.

‡jpenalva@emp.uc3m.es, Universidad Carlos III de Madrid and Banco de España.

Keywords: High frequency traders, high frequency trading, flash trading, liquidity traders, institutional investors, market microstructure

1 Introduction

Around 1970, Carver Mead coined the term "Moore's law" in reference to Moore's statement that transistor counts would double every year. There is some debate over whether this "law" is empirically valid but there is no discussion that the last forty years have seen an explosive growth in the power and performance of computers. Financial markets have not been immune to this technological advance, it may even be one of the places where the limits of computing power are tested every day. This computing power is harnessed to spot trends and exploit profit opportunities in and across financial markets. Its influence is so large that it has given rise to a new class of trading strategies sometimes called algorithmic trading and others high frequency trading. We prefer to use algorithmic trading (AT) as the generic term that refers to strategies that use computers to automate trading decisions, and restrict the term high frequency (HF) trading to refer to the subset of AT trading strategies that are characterized by their reliance on speed differences relative to other traders to make profits and also by the objective to hold essentially no asset inventories for more than a very short period of time.[1]

The advent of AT has changed the trading landscape and the impact of their activities is at the core of many regulatory and financial discussions. The explosion in volume of transactions we have witnessed in the last decade, and the speed at which trades are taking place, is highly suggestive that AT is very much in use and that these strategies are not being driven out of the market as a result of losses in their trading activities. Indeed, different sources estimate that annual profits from AT trading are between $3 and $21 billion (Brogaard [2010] and Kearns et al. [2010]). These strategies have supporters and detractors: on one side we find trading houses and hedge funds who vigorously defend their great social value, whilst being elusive about the profits they make from their use; and on the other hand there are trading

[1]This definition is consistent with the one used in Kyle [Flash Crash p3]: "We find that on May 6, the 16 trading accounts that we classify as HFTs traded over 1,455,000 contracts, accounting for almost a third of total trading volume on that day. Yet, net holdings of HFTs fluctuated around zero so rapidly that they rarely held more than 3,000 contracts long or short on that day."

houses that denounce high frequency traders (HFTs) as a threat to the financial system (and their bottom line). Although AT in general and HF trading in particular have been in the market supervisors' spotlight for quite some time and efforts to understand the consequences of HF trading have stepped up since the 'Flash Crash' in May 6 2010 (SEC [2010], Commission et al. [2010], Kirilenko et al. [2010], and Easley et al. [2011]) there is little academic work that addresses the role of these trading strategies. The objective of this paper is to provide a framework with which to analyze the issues surrounding HF trading, their widespread use, and their value to different market participants.

To analyze the impact of HF trading in financial markets we develop a model with three types of traders: liquidity traders (LTs), market makers (MMs), and HFTs. In this model LTs experience a liquidity shock and come to the market to unwind their positions which are temporarily held by the MMs in exchange for a liquidity discount. HFTs mediate between LTs and MMs. HFT mediation instantaneous, buying from one and selling to the other while holding no inventory over time. HFTs, because of their information processing and execution speed, make profits from this intermediation by extracting trading surplus. The same model without HFTs, which corresponds to that of Grossman and Miller [1988], serves as benchmark to analyze the impact of HFTs.

Naturally, we find that HFT's additional intermediation increases the volume of trade substantially (it doubles). The additional volume is neither driven by fundamentals (only the original trades, without the HFT, are driven by fundamentals) nor is it noise trading. Far from it, the extra volume is a consequence of trades which are carefully tailored for surplus extraction. Moreover, HF trading strategies introduce "microstructure noise": in order to profit from intermediation HFTs buy shares from one trader at a cheap price and sell it more dearly to another trader, generating price dispersion where before there was only a single price. These properties, which are built into the model, closely correspond to observed behavior (e.g. Kirilenko et al. [2010]).

Our main findings are: (i) the presence of HFTs exacerbates the price impact of the initial liquidity trades that generate a temporary order imbalance, imposing a double burden on liquidity demanders: the direct cost from the trading surplus extracted by the HFT, and the

indirect cost of a greater price impact; (ii) furthermore, this effect is increasing in the size of the liquidity need, consistent with the results in Zhang [2010]; (iii) the higher initial price impact arises as traders anticipate the future additional trading costs from the presence of HFTs, which generates an increase in the liquidity discount. Thus, MMs suffer countervailing effects from HFTs: increased trading costs from HFT surplus extraction versus increased expected returns from higher liquidity discounts. In our model these two effects cancel each other, leaving expected profits for MMs unchanged. (iv) Standard measures of market liquidity may lead to erroneous conclusions: in our model, HFTs do not increase liquidity and yet we observe increased trading volumes. In fact, liquidity traders face overall lower sales revenue and higher costs of purchase, suggesting that liquidity is better measured through total cost of trade execution. (v) Finally, we consider competition between HFTs and the decision for MMs to become HFTs. We find that profits from HF trading attract entrants who are willing to invest in acquiring the skills necessary to compete for these profits, but that competition is limited. As the number of HFTs increases, the expected profits of HF trading falls until the expected skills of an entrant (relative to those of existing HFTs) are insufficient to generate enough profits to cover the initial investments required to become an HFT. Thus, in equilibrium traditional MMs with low expected skills as HFTs will continue in their traditional role, coexisting with others acting as skilled and profitable HFTs.

Our analysis focuses on the effect of HFT's surplus extraction on trades initiated by liquidity needs that generate temporary trading imbalances. Nevertheless, our analysis of the role and effect of HFTs also applies to a broader set of circumstances. In particular it applies to trading by mutual fund managers, hedge funds, insurance companies and other large investors, and trading motivated not only by immediate liquidity needs, but also trading to build up or unwind an asset position, for hedging, etc.

Two contemporaneous empirical papers lend strong support to the stylized features that our theoretical model captures as well as the implications concerning the impact that HF trading has on financial markets. The recent work of Zhang [2010] firmly concludes that HF trading increases stock price volatility and that this positive correlation between volatility and HF trading "is stronger for stocks with high institutional holdings, a result consistent with the view

that high-frequency traders often take advantage of large trades by institutional investors". Kirilenko et al. [2010] study the impact of HF trading during the Flash Crash on May 6 2010. Their findings about the activities of HFTs also provide strong support for the theoretical description we use to include HFTs as pure surplus extractors in our theoretical model. They find that HFTs have among all types of traders the highest price impact and that "HFTs are able to buy right as the prices are about to increase. HFTs then turn around and begin selling 10 to 20 seconds after a price increase." Moreover, they find that "The Intermediaries sell when the immediate prices are rising, and buy if the prices 3-9 seconds before were rising. These regression results suggest that, possibly due to their slower speed or inability to anticipate possible changes in prices, Intermediaries buy when the prices are already falling and sell when the prices are already rising." These findings strongly support our assumption that HFTs (due to their speed advantage) can for the most part effectively anticipate and react to price changes as a key part in their strategies for surplus extraction.

Before delving into our analysis of HF trading, in Section 2 we provide a brief overview of HF trading and HFTs, what HFTs could be doing, and what is it about trading speed that is so profitable for some and damaging for others. After this quick overview, in Sections 3 and 4 we develop our framework and analysis with a single HFT, and use the model to discuss the main issues raised by the presence of HFTs, respectively. In Section 5 we introduce competition amongst HFTs and the decision of an MM who considers setting up an HF trading desk. In Section 6 we conclude and discuss some key features about HFTs that require further research (and quality data).

2 Trading Algorithms, High Frequency Traders, and Financial Markets

2.1 Financial Market Developments

Over the last years all major exchanges have revamped their systems to give way to the new era of computerized trading. Speed of trading and volume figures speak for themselves. In

the SEC's report on "Findings regarding the market events of may 6, 2010" (SEC [2010]) we read that NYSE's average speed of execution for small, immediately executable orders was 10.1 seconds in January 2005, compared to 0.7 seconds in October 2009. Also, consolidated average daily share volume in NYSE-listed stocks was 2.1 billion shares in 2005, compared to 5.9 billion shares in January through October 2009. Consolidated average daily trades in NYSE-listed stocks was 2.9 million trades in 2005, compared to 22.1 million trades in January through October 2009. Consolidated average trade size in NYSE-listed stocks was 724 shares in 2005, compared to 268 shares in January through October 2009.

Other important metrics that intend to capture market efficiency and information transmission have also undergone considerable changes as a result of modifications of market rules and the prominent role that computing has taken in financial markets. For example, Chordia et al. [2010] focus on comparisons of pre- and post-decimal trading in NYSE-listed stocks (subperiods from 1993-2000 and 2001-2008). Some of their findings are that average effective spreads decreased significantly (from \$0.1022 to \$0.0223 cents for small trades (<\$10,000) and from \$0.1069 to \$0.0267 for large trades (>\$10,000)), while average depth available at the inside bid and offer declined significantly (from 11,130 shares to 2,797 shares). From 1993-2000 the mean trade size is \$82,900 and from 2001-2008 \$36,400 while the mean number of transactions is 1,136 and 14,779 respectively.

2.2 What characterizes Algorithmic and HF Trading

We adopt Hendershott et al. [2010]'s definition of AT: "the use of computer algorithms to automatically make certain trading decisions, submit orders, and manage those orders after submission". We distinguish HF trading as a subset of AT. An HF trading strategy is an AT strategy that is based on exploiting greater processing and execution speed to obtain trading profits while holding essentially no asset inventory over a very short time span—usually measured in seconds, mostly less than a few minutes, and certainly less than a day.[2] One sometimes finds these strategies described as *latency arbitrage*. HFTs are proprietary firms and

[2]In their study of the 'Flash Crash', Kirilenko et al. [2010] find that, holding prices constant, HFTs reduce half their net holdings in 115 seconds.

proprietary trading desks in investment banks, hedge funds, etc, that based on these strategies have the ability to generate large amounts of trades over short periods of time, Cvitanić and Kirilenko [2010]. There are other AT strategies that are in use for other purposes. For example, there are AT liquidity strategies which strategically post and cancel orders in the order book to exploit widening spreads, or AT strategies designed to execute large orders with the smallest price impact. Our analysis focuses exclusively on HF trading strategies, which we believe are the ones most critics of AT have in mind.

Making the distinction between HF trading and AT is important because it highlights the substantial difficulty one encounters when measuring the impact that HFTs have on markets according to metrics such as volume, spreads, and liquidity. For example, estimates of HF trading volume in equity markets vary widely depending on the year or how they are calculated, but they are typically between 50% and 77% of total volume, see SEC [2010] and Brogaard [2010]—although how much is actual HF trading versus generic AT is unclear. Also, Hendershott et al. [2010] find that for large-cap stocks AT improves liquidity and narrows effective spreads. They also find that AT increases realized spreads which indicates that revenue to liquidity suppliers has increased with AT, but it is difficult to infer how much of these effects are due to AT that is not HF trading. Similar identification problems are present in another recent study, Brogaard [2010], which finds that HFTs contribute to price discovery and reduce volatility. Thus, this identification problem, as well as possible collateral effects on other AT strategies, have to be taken into account in any regulatory implications one may draw from our analysis, as we focus exclusively on HF trading.

As per our definition, paramount to the activities of HF traders is the speed at which they can: access and process market information; generate, route, cancel, and execute orders; and, position orders at the front of the queue in the trading book to avoid having stale quotes in the market. Their speed or low latency is mainly due to two key ingredients: capacity (software and hardware), and co-location. Co-location allows HFTs to place their servers in close physical proximity to the matching engines of the exchanges. Surprisingly, being near the exchanges can shave the speed of reaction by a sufficient number of fractions of a second to provide HFTs

a valuable edge when trading in the market—to the extent that they are willing to pay millions of dollars for this service.

Perhaps the most revealing behavior of HFTs is how they make use of cancelations to poke the market and extract valuable information. For instance, the strategy known as 'pinging' is based on submitting immediate-or-cancel orders which are used by HFTs to search for and access all types of undisplayed liquidity SEC [2010].[3] Another strategy, known as 'spoofing', consists of sending out a large amount of orders over a short period of time before immediately canceling most of them so that only a few are executed. This burst of activity is expected to trigger other algorithms to join the race and start buying or selling (and slow down information flows to other market participants).

2.3 How is it that HFTs could be making money?

Here we provide four stylized examples that show how HFTs could be exploiting their speed advantage by posting, executing, and canceling orders to position their orders at the front of the queue and intermediate in market transactions for a negligible period of time. Although the first three examples outline different strategies used by HFTs, the underlying feature common to all three is the ability that HFTs have to extract trading surplus.

Example 1. One way in which HFTs can extract surplus is by exploiting their speed to alter market conditions in a way that encourages buyers to accept a slightly higher price and sellers a slightly lower one. The idea is relatively simple and works in a setting where liquidity traders split their trades in small packages and MMs do not have large outstanding offers in the books. Suppose a trader (LT) needs liquidity and wants to sell a block of shares. As the first shares come into the system (say at the best buy price of \$5.50 per share), the HFT cancels her outstanding posted buy offers that have not been executed. She then posts additional sell offers, adding to the increased selling pressure, in order to help clear the remaining posted buy orders in the book. Once the book is clear, the HFT quickly reposts a significant number of offers at lower prices (say \$5.47) so that she is first in the buying queue. This is only possible if

[3]There are circumstances when orders are placed close to best buy or sell with no intention to trade, this is known as book layering, and up to 90% of these orders are immediately canceled, see SEC [2010].

she can move quickly enough that by the time the MM reacts to the increased selling pressure, new posted offers by the MM sit behind those posted by the HFT at $5.47 per share. The LT finds that the market around $5.50 has dried up and can only sell at $5.47. These shares are bought by the HFT, who is at the front of the queue. Also, having posted substantial orders at the front of the queue at $5.47 allows the HFT to be the first to notice when the liquidity pressure eases. At that moment, the HFT cancels her posted buy offers and starts posting sell offers at slightly higher prices (say $5.49 first and then at $5.48). The MM sees the selling pressure shift and the price rebound. Although the MM is sorry to see the lower prices disappear, he is still willing to buy at $5.49, which allows the HFT to sell her earlier purchases and end up with a zero net position. During this process, the HFT has been able to make profits on the shares bought at $5.47 and sold at $5.48 or $5.49, while taking a loss on the shares executed at the beginning, bought at $5.50 and sold at $5.48 or $5.49. An HFT who is fast enough will have bought many more shares at $5.47 than at $5.50 by selectively canceling and reposting her offers to make the strategy profitable. As for the other traders: The MM bought some shares below the initial best buy price so he is satisfied; the other LT is also satisfied from having been able to sell his shares even though for some of them he received a cent less than what the MM paid for them.

Example 2. Assume that the market opens after the release of good news about a company's performance. At opening, shares are selling at $5.50 and slow traders (traders who are not HFT) are posting buy orders. Due to the high latency of the slow traders, HFTs see the buy orders arriving in the system and decide to purchase as many shares as possible at the current price. Immediately, HFTs issue low volume immediate-or-cancel sell orders to gauge how much are slow traders willing to pay for the shares. For example, an immediate-or-cancel sell order goes out at a price of $6.00 per share and if it does not get filled it is immediately canceled and a new immediate-or-cancel sell order is sent at $5.99 and so on until it is filled, say at $5.70. At this point the HFTs unload all shares that were purchased making a profit and holding no inventories.

Example 3. Similarly, HFTs may take advantage of the so-called flash orders which give them an informational edge over other traders in the market. For example, a buy order for

1,000 shares at \$5.50 is 'flashed' to a reduced number of traders some of which are HFTs. Traders that are flashed the information not only know about such a potential trade before the vast majority of the market, but they are also capable of acting upon the information before it reaches the rest of the market. If the HFTs believe or are able to correctly anticipate that this buy order is part of a large lot that will trickle through the system in small batches of, say 1,000 shares at a time, they have time to react and purchase all shares in the market; use sell-or-cancel orders to find out the upper limit at which the counterparty that initiated the buy order is willing to pay to liquidate his entire lot of shares; and complete the round trip of buying and selling whilst making a profit and carrying no inventory, all of this in the intraday market, and possibly within a couple of minutes.

Example 4. Other trades that could be profitable for HFTs, but do not profit from extracting surplus, are those that are designed to collect rebates offered by the exchange. Market centers attract volume by offering a liquidity rebate to MMs that post orders. Rebate strategies that break even by selling (buying) and then buying (selling) shares at the same price are profitable because they earn the rebate for providing liquidity. If the rebate is around 0.25 cents per share then a round trip rebate accrues half a cent per share to the HFT that pursues this type of rebate trading.

We note that although the examples above have been contrived so that the HFTs make positive profits these strategies are not arbitrages. This is because there are states of the world where they can also deliver a loss. These strategies are not riskless and the HFTs face different types of risk, for instance volume and price risk. In all cases the HFT must take a view on the direction the market is taking and how many shares and at what prices she is willing to buy (sell) before turning around to sell (buy) her holdings with the objective of not carrying inventories for too long. Here we take the view (supported by the findings in Kirilenko et al. [2010]) that as a result of their speed of execution, but more importantly, of information processing, together with their ability to post and cancel orders, it is presumed that more often than not the HFT will earn a profit, or break even, and be flat after a short period of time. It is the value of these trades we want to analyze.

3 The Model

We study the role of HF trading in a context where the stock market has social value because it facilitates the financing of economic activity, adding value to equity holders by providing a way for them to convert their equity into cash (and viceversa) quickly and at a reasonable price. Grossman and Miller [1988] (GM) provide a highly stylized model of such a stock market. In the GM model, equity investors (liquidity traders) quickly find counterparties for their trading needs.[4] These counterparties are MMs who are willing to take the other side of investor trades and hold those assets temporarily–until another investor enters the market to eliminate the temporary order imbalance.[5] Holding these assets entails price risk, which MMs are willing to bear in exchange for a small discount on the asset's price (the liquidity discount). We introduce HFTs in this model, where an HFT is a trader who, thanks to her rapid information processing and quick execution ability, can extract part of the trading surplus from the transactions between equity investors and MMs.[6]

The setting for our model is a world with three dates, $t = 1, 2, 3$ in which a temporary order imbalance of size i affects conditions in a stock market with a cash asset (with a return normalized to zero) and a risky asset. There are two "outside investors", which we refer to as liquidity traders (LT1 and LT2). At time 1, LT1 is endowed with cash and an amount i of shares which due to a liquidity shock he would like to sell (that is, he wants to trade $-i$ shares). On the other hand, the other trader, LT2, wants to buy $i > 0$ units (trade i shares) but only reaches the market at date 2.[7] If both traders met at the same time they would exchange the asset at the current "fundamental" price, which is the expected future cash value

[4]The other interpretation of the model is intended for the futures market. Everything we say for the stock market is also equally valid for trading in the futures market in exactly the same terms. We refer the interested reader to the original article for details of how to reinterpret the model.

[5]Although the GM model is designed for specialist markets, our analysis can be extrapolated to markets with an open limit order book, in which there is a cost for monitoring the order book. In such markets, the role of MMs is represented by traders who are permanently monitoring the order book and who, despite having no privileged information, are willing to assume trading positions for a possibly significant period of time

[6]In the exposition we assume that the HFT extracts trading surplus with certainty to reduce notation. If we assume that the HFT extracts surplus with a given probability, $q \in (0, 1]$, and with probability $p \in [0, 1 - q]$ makes a loss, the qualitative nature of the results continues to hold as long as the HFT obtains positive expected profits from trade surplus extraction.

[7]We make the assumption that LT1 wants to sell and LT2 wants to buy to streamline the presentation. The analysis is equally valid if LT1 wants to buy i shares and LT2 wants to sell them.

of the asset. But, as LT1 arrives to the market much earlier than LT2, his only option is to sell to the M intermediaries, the MMs. The intermediaries will accept LT1's order at date 1 at the "right price" and hold the shares until LT2 enters the market at date 2. The "right price", which we call the *market price* of the asset, is determined by the fundamental price minus a discount commanded by the MMs for holding the asset until date 2. MMs bear the risk of having to sell the asset at a loss after the release of public news that might adversely change the fundamental price of the asset, and they require the discount to compensate them for this risk. This "liquidity discount" is calculated as the difference between the "fundamental" value of the asset and the price at which the transaction takes place, the market price.

All these traders (LT1, LT2, and MMs) are price-taking and risk averse. They have expected utility from wealth at date 3 and they choose their asset positions so as to maximize $\mathbb{E}\left[U\left(W_3\right)\right]$, where $U\left(W\right) = -\exp\left(-aW\right)$ is the utility function, W is wealth and a is the risk aversion parameter. The future cash value of the (non-dividend paying) asset is $P_3 = \mu+\epsilon_2+\epsilon_3$, where μ is constant, and ϵ_2 and ϵ_3 are normally and independently distributed with mean 0 and variance σ^2. The random variables, ϵ_2 and ϵ_3, represent public information that is announced between dates $t = 1$ and $t = 2$, and between $t = 2$ and $t = 3$ respectively. Let $\mu_2 = \mathbb{E}\left[P_3|\mathcal{F}_2\right] = \mu + \epsilon_2$ denote the expectation of P_3 conditional on information available at $t = 2$. We use the notation $\theta_t^x\left(P_t\right)$ to describe the asset holdings of trader x after trading at date t given price P_t. A positive asset position, $\theta_t^x > 0$, implies holding (long) a positive number of shares.

3.1 HF trading and the Extraction of Trading Surplus

We start with a single, monopolistic HFT in the market. Below, in Section 5, we discuss competition between HFTs. The monopolistic HFT exploits her greater processing and execution speed to extract trading surplus from transactions at each date. In our highly stylized setting, this is modeled by allowing the HFT to take over any trader's position in two blocks: the first block is executed at the market price P_t plus or minus a haircut Δ (depending on whether the HFT is buying or selling) and the other at the market price. The profits the HFT can obtain from this ability depend on market conditions and is illustrated by the following simplified example which is followed below by the description of the general model.

Suppose that a liquidity trader (LT) wants to sell $1,000$ shares, the expected price of the asset is \$5.50 and there are nine MMs. Suppose further that all traders are risk averse and they have linear demands for assets given by

$$\theta^{MM}(P) = \theta^{LT}(P) = 5000(5.50 - P) .$$

Hence, they will only buy shares if the price is below \$5.50.

In a trading equilibrium, the holdings of all traders (one LT and nine MMs) have to add up to the supply of LT (1000 shares) so that

$$\begin{aligned} 9\theta^{MM}(P) + \theta^{LT}(P) &= 1000 \\ \Rightarrow 9 \times 5000(5.50 - P) + 5000(5.50 - P) &= 1000 \\ \iff P &= 5.48 . \end{aligned}$$

At the price of \$5.48 per share, each trader (including the LT) will hold 100 shares. This implies that the (price sensitive) LT who initially wanted to sell 1,000 shares (at \$5.50) ends up only selling 900 shares at a price of \$5.48 (the fundamental price \$5.50 minus a 2 cent liquidity discount).

The HFT can enter the market and make profits by introducing a haircut, say of 1 cent per share, inducing the LT to sell part of his shares at a price of \$5.47 and inducing MMs to buy part of the package at the slightly higher price of 5.49 (using one of the strategies discussed in Section 2.3).

The HFT's profits are determined by traders' asset demands. At a price of \$5.47, LT will hold $5000(5.50 - 5.47) = 150$ shares, thus selling only 850 shares. At a price of \$5.49, MMs will hold $5000(5.50 - 5.49) = 50$ shares each, that is 450 shares. These do not cancel and implies the HFT would be left holding 400 shares. This is not satisfactory for the HFT who wants to hold zero inventory at the end of the trading round. Thus, she modifies her trading strategy slightly: After offering to buy 850 shares from LT at 5.47, she offers to buy another batch of 50 shares from the LT at \$5.48. The LT finds these trades acceptable and executes them.

Having acquired all LT's shares, the HFT turns around and first offers to sell 50 shares to each of the MMs at \$5.49. Then, she offers to sell them another 50 shares each at \$5.48. Again, this combination of offers is acceptable to the MMs and they agree to execute them. The final result is that the HFT has made 2 cents on each of 450 shares sold at \$5.49 (and bought at \$5.47) and another cent on 400 shares bought at \$5.47 and sold at \$5.48 for a total of \$13.00. Furthermore, she ends up holding no inventory. The market has seen a doubling of the trading volume, plus some transactions taking place at prices above and below the "market price" of the asset (\$5.48)–the fundamental price (\$5.50) minus the equilibrium liquidity discount (\$0.02).

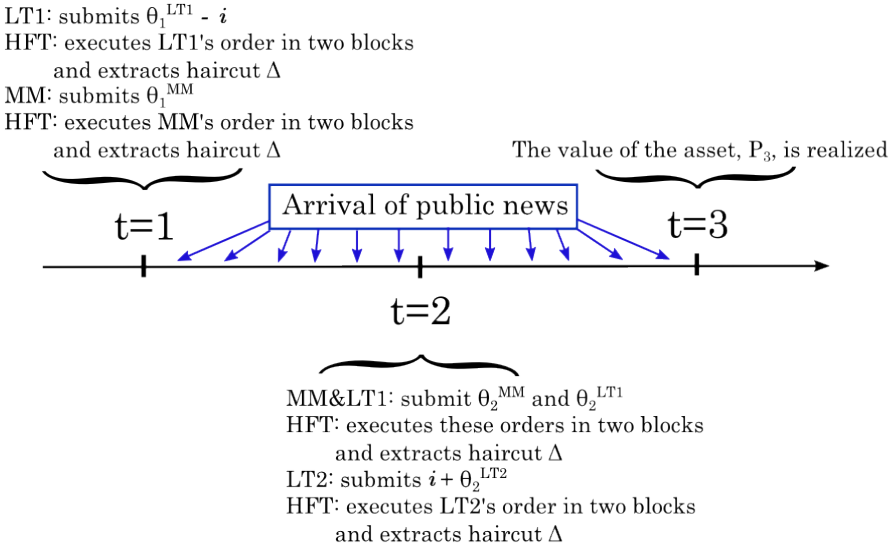
This example represents what happens in the extended general model from which we can derive equilibrium prices, haircuts, volumes and the effect of the HFT on the number of MMs in the economy. Although the example (and the model) is highly stylized—it exaggerates the profits that an HFT could extract from the transaction, and involves knowledge that is not directly available to any trader—it does capture the advantage gained by HFTs through their speed when it comes to extracting and processing information from the order flow and using their execution speed to cancel and advantageously repost trades in the order book, as was discussed above.

3.2 The general model

The timeline is as follows (Figure 1): at date 1, the price sensitive trader LT1 comes to the market wanting to sell i shares. He is price sensitive so that his date 1 excess demand is given by $\theta_1^{LT1}(P_1) - i$. The HFT trades with LT1 in two blocks, he buys the first block at $P_1 - \Delta$ and the other at P_1. Immediately, and still at date 1, the HFT turns around and sells to the MMs in two blocks: one block at $P_1 + \Delta$ and the other at P_1. As time passes between dates 1 and 2, there are public announcements of news about the future value of the asset, P_3, which generates fluctuations in the expected value of the asset. At date 2, the other price-sensitive liquidity trader LT2 decides to enter the market to acquire i shares. His excess demand is $\theta_2^{LT2}(P_2) + i$. The HFT quickly purchases all the shares that the MMs had bought in date 1, plus those that LT1 could not sell in date 1, in two blocks: one block at $P_2 - \Delta$ and the other

at P_2. Then the HFT turns around and sells all the shares to LT2 in two blocks: one at $P_2+\Delta$ and the other at P_2.[8]

Figure 1: Timeline of trading in the presence of the HFT



We now proceed by solving the general model by backward induction, starting at date 2 and then moving one period back to the problem at date 1.

Trading at $t=2$

The second liquidity trader, LT2, enters at $t=2$ and given price P_2, this price sensitive trader's excess demand is $\theta_2^{LT2}(P_2)+i$ shares (where we recall that $\theta_2^{LT2}(P_2)$ is the asset holding after trading at date 2), but he knows that there is an HFT who will extract some surplus. In our model, any trader who wants to buy $\theta(P)+i$ units at price P will have to go through the HFT who will split the order in two blocks at quantity $\tilde{\theta}$: the HFT will sell $\tilde{\theta}+i$ units at price $P+\Delta$ and the remaining $\theta(P)-\tilde{\theta}$ units at P. To ensure trades are mutually acceptable at all prices the quantity $\tilde{\theta}$ corresponds to the trader's excess demand at $P+\Delta$, that is $\tilde{\theta}=\theta(P+\Delta)$. We assume traders accept trading in two blocks and paying the haircut as part of the cost of doing

[8]The time sequence of trades at $t=2$ is not important, as the HFT can sell short shares to LT2 and then cover his short positions by trading with MMs.

business, and that they take the quantity $\tilde{\theta}$ as exogenously given.[9] We use the notation $\tilde{\theta}_t^x$ to describe the quantity that the HFT uses to split trader x's desired trades at date t into the two blocks. The consistency condition, $\tilde{\theta} = \theta(P + \Delta)$, coupled with the fact that the HFT splits the orders into only two blocks, limits the advantage gained by the HFT.

We also limit the advantage gained by the HFT in terms of the information she can use to set haircuts. In principle, the HFT could distinguish between different traders and different periods of time. But, being able to make such fine grained distinctions between traders would require extracting an amount of information from the order flow that we consider unrepresentative of what goes on in stock markets today. Nevertheless, we do believe that HFTs can extract sufficient information from the order flow at a sufficient speed to allow them to distinguish a large liquidity trade coming in (one that will move the market price substantially) from regular trades posted by other market participants. We capture this by allowing the HFT to distinguish "large" trades (the initial trade by LT1 at date 1 and LT2's trade at date 2) from "small" trades (trades by MMs and LT1 at $t = 2$). Thus, the HFT sets two haircuts, one for large trades Δ_L, and one for small trades Δ_S.[10]

Since LT2's excess demand is $\theta_2^{LT2}(P_2) + i$, the HFT applies the large haircut to him and sells $i + \tilde{\theta}_2^{LT2}$ shares at the (higher) price, $P_2 + \Delta_L$, and $\theta_2^{LT2}(P_2) - \tilde{\theta}_2^{LT2}$, at the market price P_2. Thus, given an initial wealth of W_2^{LT2}, LT2's final wealth is

$$W_3^{LT2} = W_2^{LT2} + (P_3 - P_2)\left(\theta_2^{LT2}(P_2) + i\right) - P_3 i - \left(\tilde{\theta}_2^{LT2} + i\right)\Delta_L,$$

which, if we drop the superscripts and the functional dependence of asset holding on prices, can be written as

$$W_3^{LT2} = W_2^{LT2} + (P_3 - P_2)\left(\theta_2^{LT2} + i\right) - P_3 i - \left(\tilde{\theta}_2^{LT2} + i\right)\Delta_L.$$

[9] It is possible to introduce a strategic element whereby the trader alters his bidding behavior in anticipation of the HFT. This can greatly complicate the model. Basically, it affects how much of the trading surplus is extracted by the HFT and imposes an additional distortion on trade.

[10] In the Appendix we have included more general formulas and additional results with slightly different assumptions on how the HFT applies haircuts across dates and traders.

Substituting for wealth in the utility function, we obtain:

$$\mathbb{E}\left[U\left(W_3\right)|\mathcal{F}_2\right] = -\exp\left(-a\left(\left(\theta_2^{LT2}+i\right)\left(\mu_2-P_2\right)-i\mu_2-W_2^{LT2}-\left(\tilde{\theta}_2^{LT2}+i\right)\Delta_L\right)+\frac{1}{2}a^2\sigma^2\left(\theta_2\right)^2\right). \tag{1}$$

LT2's final excess demand, $\theta_2^{LT2}+i$, is such that it maximizes (1). Hence,

$$\theta_2^{LT2}+i = \frac{\mu_2-P_2}{a\sigma^2}+i\,. \tag{2}$$

Similarly, LT1's and MMs' excess asset demands at $t=2$ are given by

$$\theta_2^{LT1}-i = \frac{\mu_2-P_2}{a\sigma^2}-i \qquad \text{and} \qquad \theta_2^{MM} = \frac{\mu_2-P_2}{a\sigma^2}\,.$$

Market clearing requires that the excess demand of LT1 (who arrived at date 1), plus MMs', plus LT2's sum to zero:

$$\theta_2^{LT1}-i+M\theta_2^{MM}+\theta_2^{LT2}+i=0\,,$$

so that the equilibrium price is $P_2=\mathbb{E}\left[P_3|\mathcal{F}_2\right]=\mu_2$, and the asset holdings after trade in date 2 are $\theta_2^{LT2}=\theta_2^{LT1}=\theta_2^{MM}=0$. The quantities used by the HFT for extracting surplus are obtained from the corresponding demand functions $\tilde{\theta}_2^{LT1}=\tilde{\theta}_2^{MM}=\theta_2^{LT1}\left(P-\Delta_S\right)$ and $\tilde{\theta}_2^{LT2}=\theta_2^{LT2}\left(P+\Delta_L\right)$:

$$\tilde{\theta}_2^{LT1}=\frac{\Delta_S}{a\sigma^2}\,, \qquad \tilde{\theta}_2^{M}=\frac{\Delta_S}{a\sigma^2}\,, \qquad \tilde{\theta}_2^{LT2}=-\frac{\Delta_L}{a\sigma^2}\,.$$

The HFT will extract trading surplus equal to[11]

$$\Pi_2\left(i\right)=\left|\tilde{\theta}_2^{LT2}+i\right|\Delta_L+\left|\tilde{\theta}_2^{LT1}-\theta_1^{LT1}\right|\Delta_S+M\left|\tilde{\theta}_2^{MM}-\theta_1^{MM}\right|\Delta_S. \tag{3}$$

[11] We are assuming that the MMs and LT1 are net sellers of the asset. This is confirmed as in equilibrium MMs will be net buyers in the first period and have zero net final holdings (and $\theta_2^{LT1}-\theta_1^{LT1}$ has the same sign as MMs' changes in asset holdings at date 2).

The general model: Trading at $t = 1$

Traders at date $t = 1$ anticipate what will happen at $t = 2$. They also know that there will be public information revealed prior to date 2 trading and that $P_2 = \mu_2 = \mu + \epsilon_2$ is random and normally distributed with mean μ and variance σ^2.

Traders' future wealth can be written as

$$W_3^{LT1} = W_0^{LT1} + (P_2 - P_1)\left(\theta_1^{LT1} - i\right) + (P_3 - P_2)\left(\theta_2^{LT1} - i\right) + P_3 i - \left(\theta_1^{LT1} - \tilde{\theta}_2^{LT1}\right)\Delta_S - \left(i - \tilde{\theta}_1^{LT1}\right)\Delta_L$$

and

$$W_3^{MM} = W_0^{MM} + (P_2 - P_1)\,\theta_1^{MM} + (P_3 - P_2)\,\theta_2^{MM} - \left(\theta_1^{MM} - \tilde{\theta}_2^{MM}\right)\Delta_S - \tilde{\theta}_1^{MM}\Delta_S\,,$$

where $\theta_t^{LT1} - i$ and θ_t^{MM} are LT1's and MM's excess demands respectively in dates 1 and 2.

Using $\mathbb{E}\left[P_2|\mathcal{F}_1\right] = \mathbb{E}\left[\mu + \epsilon_2\right] = \mu$, simplifies the expression for traders' wealth, and it is straightforward to derive optimal excess demands:

$$\theta_1^{LT1} - i = \frac{1}{a\sigma^2}\left(\mu - P_1 - \Delta_S\right) - i \qquad \text{and} \qquad \theta_1^{MM} = \frac{1}{a\sigma^2}\left(\mu - P_1 - \Delta_S\right).$$

Notice that traders anticipate that their current asset demand (and hence positions at the end of trading at $t = 1$) will affect future trading and hence the (date 2) haircuts they will have to pay. The date 1 market clearing condition is

$$\left(\theta_1^{LT1} - i\right) + M\theta_1^{MM} = 0\,.$$

From this we obtain: (i) the market clearing price

$$P_1 = \mu - \Delta_S - \frac{ia\sigma^2}{M+1}\,, \tag{4}$$

(ii) traders' asset holdings after trading in date 1,

$$\theta_1^{LT1} = \frac{i}{M+1} \qquad \text{and} \qquad \theta_1^{MM} = \frac{i}{M+1}\,,$$

and, using $\tilde{\theta}_1^{LT1} = \theta_1^{LT1}(P - \Delta_L)$ and $\tilde{\theta}_1^{MM} = \theta_1^{MM}(P + \Delta_S)$, (iii) the quantities that are subject to haircuts

$$i - \tilde{\theta}_1^{LT1} = i - \frac{i}{M+1} - \frac{\Delta_S}{a\sigma^2} \qquad \text{and} \qquad \tilde{\theta}_1^{MM} = \frac{i}{M+1} - \frac{\Delta_S}{a\sigma^2}\,.$$

Again, the HFT is able to extract trading surplus and generate profits of:

$$\Pi_1(i) = \left|i - \tilde{\theta}_1^{LT1}\right| \Delta_L + M \left|\tilde{\theta}_1^{MM}\right| \Delta_S. \tag{5}$$

In summary, from the above analysis we extract the following conclusions:

Theorem 3.1. *For a given order imbalance of magnitude $i > 0$:*

1. *Market clearing prices are*

$$\begin{aligned} P_1 &= \mathbb{E}[P_2|\mathcal{F}_1] - \frac{ia\sigma^2}{M+1} - \Delta_S\,, \\ P_2 &= \mathbb{E}[P_3|\mathcal{F}_2] = \mu_2\,, \end{aligned} \tag{6}$$

and the liquidity discount at $t = 1$ is

$$\mathbb{E}[P_2|\mathcal{F}_1] - P_1 = \frac{ia\sigma^2}{M+1} + \Delta_S\,. \tag{7}$$

2. *Asset trading is described on Table 1:*

Table 1: Prices and Volume of Trades

Price	LT1	LT2	MM (total)
$P_1 - \Delta_L$	$-\frac{M}{M+1}i + \frac{\Delta_L}{a\sigma^2}$		
P_1	$-\frac{\Delta_L}{a\sigma^2}$		$M\frac{\Delta_S}{a\sigma^2}$
$P_1 + \Delta_S$			$\frac{M}{M+1}i - M\frac{\Delta_S}{a\sigma^2}$
$P_2 + \Delta_L$		$i - \frac{\Delta_L}{a\sigma^2}$	
P_2	$-\frac{\Delta_S}{a\sigma^2}$	$\frac{\Delta_L}{a\sigma^2}$	$-M\frac{\Delta_S}{a\sigma^2}$
$P_2 - \Delta_S$	$-\frac{1}{M+1}i + \frac{\Delta_S}{a\sigma^2}$		$-\frac{M}{M+1}i + M\frac{\Delta_S}{a\sigma^2}$

3. The HFT acts as counterparty to all these trades and makes profits equal to

$$\Pi(i) = \Pi_1(i) + \Pi_2(i) \ , \tag{8}$$

where $\Pi_1(i)$ *and* $\Pi_2(i)$ *are described in Equations (5) and (3) respectively.*

The presence of an HFT affects the market clearing price when the order imbalance is initiated. Traders, anticipating future haircuts imposed by the HFT, amplify the selling pressure of the liquidity trader initiating the order imbalance (LT1), driving the date 1 market clearing price down further than it would have been without HF trading (that is, with $\Delta_L = \Delta_S = 0$), and increasing the liquidity discount paid by LT1 to encourage MMs to buy the asset from him (Equation (7)).

Table 1 illustrates a number of the properties and implications of our stylized model. For example, our model exhibits "microstructure volatility" induced by the HFT. We can observe how the HFT's surplus extraction, which must by necessity involve buying and selling at different prices, introduces additional prices at which transactions take place. These price movements around the market clearing price have no informational content and may lead a casual outside observer to erroneous conclusions about what the "true" market price of the asset should be at each date.

Naturally, having allowed the HFT to intermediate all trades, the volume of trade doubles. As we will discuss later, these additional trades (and hence HFTs) do not add liquidity–in fact, the presence of HFTs results in an increase in the liquidity discount, thus liquidity is reduced. Hence, rebate schemes based on volume are likely to disproportionately benefit HFTs who are not providing liquidity, as the following simplistic rebate scheme illustrates:

Corollary 3.2. *If there is a rebate of c cents per share, the HFT gets half of all rebates, and the M market makers split a quarter of all rebates between them.*

3.3 Optimal Haircuts

The HFT, as a monopolist, realizes that there is a trade-off between a larger haircut, and a smaller number of assets traded using that haircut. Thus, she will adjust her behavior by setting haircuts that maximize the profits of her trading activity. As the HFT distinguishes the haircut for large trades (Δ_L) from the one for the small trades (Δ_S) the HFT's profits, described in Equation (9), can be expressed as:[12]

$$\Pi(i) = \Delta_L\left(\frac{M}{M+1}i - \frac{\Delta_L}{a\sigma^2}\right) + \Delta_S\left(\frac{M}{M+1}i - M\frac{\Delta_S}{a\sigma^2}\right) + \Delta_L\left(i - \frac{\Delta_L}{a\sigma^2}\right) + \Delta_S\left(i - \frac{(M+1)\Delta_S}{a\sigma^2}\right), \tag{9}$$

and by maximizing (9) with respect to Δ_L and Δ_S we obtain:

Lemma 3.3. *The HFT maximizes profits by choosing*

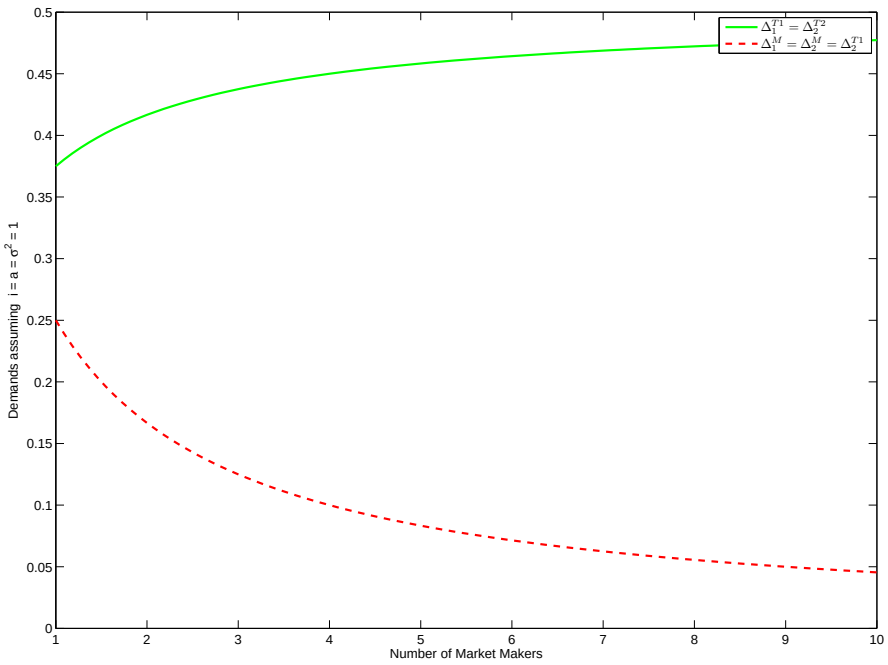
$$\begin{aligned}\Delta_S &= \frac{1}{2}\frac{1}{(M+1)}ia\sigma^2,\\ \Delta_L &= \frac{1}{4}\frac{2M+1}{M+1}ia\sigma^2 = \Delta_S\left(M+\frac{1}{2}\right).\end{aligned}$$

Haircuts are increasing in the size of the liquidity need, thus microstructure volatility and liquidity discounts will increase for larger order executions. Furthermore, since the HFT can distinguish between large and small trades, she will impose bigger haircuts on large trades than on small trades, and the increase in haircut for large trades is proportional to the number

[12] In the Appendix we discuss another case (where the HFTs make no distinctions and apply the same haircut to all trades).

of non-liquidity seeking market participants. Figure 2 below illustrates how the haircut for the large (small) trades increases (decreases) with M, and Figure 3 shows the optimal asset holdings, θ_t^x and $\tilde{\theta}_t^x$, for the liquidity traders and MMs.

Figure 2: Optimal Haircuts for large and small trades

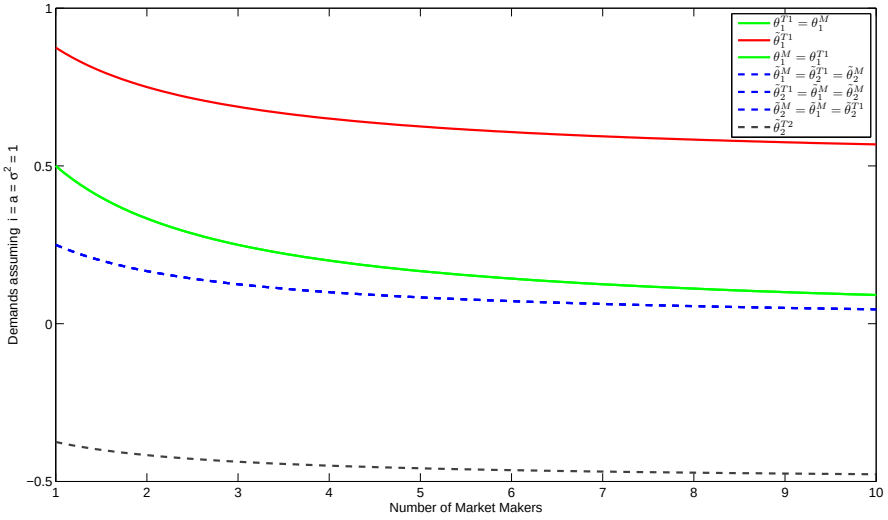


Moreover, we see that in a market with more MMs, the HFT increases the haircut imposed on large trades from liquidity seekers, LT1 at date 1 and LT2 at date 2, and lowers it on small trades. In the limit, as the number of MMs goes to infinity,

$$\lim_{M\to\infty} \Delta_S = 0 \quad \text{and} \qquad \lim_{M\to\infty} \Delta_L = \frac{1}{2} ia\sigma^2 \,,$$

we see that the haircut on small trades disappears and all the trading surplus extracted by the HFT comes from large trades. Also, as the number of MMs goes to infinity, the liquidity discount (Equation (7)) goes to zero as well, so that LT1 and LT2 behave in the same way: their liquidity needs are immediately satisfied (LT1 sells i in date 1), but they trade half their shares at a price of $\mu \pm \frac{1}{2} ia\sigma^2$ and the other half at the asset's fundamental value, μ. Then, overall the HFT then makes profits equal to $\frac{1}{2} i^2 a\sigma^2$.

Figure 3: Optimal asset holdings when the HFT sets a haircut for large trades and another haircut for small trades.



4 Measuring the impact of HFTs on financial markets

To the best of our knowledge there are only two academic articles that have measured the impact of AT in financial markets.[13] As discussed in the introduction Hendershott et al. [2010] find that AT improves liquidity and narrows effective spreads and Brogaard [2010] finds that HFTs contribute to price discovery and reduce volatility. Although the study of Brogaard [2010] has data of what he labels as HFTs, his data set cannot differentiate how much of the activity of the 26 firms in his study is AT and how much is HF trading; the same applies to the results of Hendershott et al. [2010].

In the absence of ultra-high-frequency data that provides details of HFTs trades, postings, cancelations, flash-trades, etc, our model helps understand how the presence of HFTs may be affecting different aggregate financial metrics such as volume, liquidity, price impact, and price volatility.

[13] In Section 2 we discussed the work of Kirilenko et al. [2010] which focuses on the Flash Crash and employs three days of data.

4.1 Liquidity, Volume and Price Impact

Our stylized model is designed to capture the value of a stock market as a forum where equity holders can convert their equity into cash (and viceversa) quickly and at a reasonable price. In practice it is usually hard to determine both the time and the overall cost of execution of significant blocks of shares, as share blocks are split in several (and increasingly many) parts executed separately, and the identity of traders is not available for empirical analysis.[14] Thus, most studies focus on measuring liquidity using spreads, execution speed, and fill rates, for small trades (less than 10,000 shares).

In our model, liquidity (speed and cost of execution) is essentially measured in two ways: (i) how many shares can the liquidity trader sell when he originates a trade imbalance, that is, how many shares does LT1 sell at date $t = 1$, relative to his desired total sale of i units; and (ii) what is the difference between the revenue obtained by liquidity traders when selling (the cost paid when buying) their total trading blocks ($-i$ and i) relative to executing them at the asset's "fundamental" value.

In our model we are able to identify two ways in which the HFT affects the amount of shares the liquidity trader can sell at $t = 1$: a direct and an indirect way. Although, having identified them we find that in both ways the net effect of HFTs on quantities is zero.

The direct way is the effect of the HFT's surplus extraction on trader's demand for assets (both on the shape of the demand function and on the equilibrium quantity demanded). Direct effects are limited. First, by assumption we have ruled out strategic effects on the part of traders vis-a-vis the HFT, which would be a natural source of distortions of asset demand functions and reduced trading (and hence liquidity). Second, the HFT's surplus extraction process is faultless. This, together with her extreme aversion to holding any inventory leads to no distortions in the final quantity of assets demanded: LT1 continues to carry over $\theta_1^{LT1} = i/(M+1)$ to date 2—the same amount as if there were no HFT.

[14] This is recognized by the SEC in their report [SEC concept release Jan10] "Measuring the transaction costs of institutional investors that need to trade in large size can be extremely complex. These large orders often are broken up into smaller child orders and executed in a series of transactions. Metrics that apply to small order executions may miss how well or poorly the large order traded overall." (p38-39)

The indirect way is through the effect of the HFT on the number of MMs, which in turn affects liquidity (in terms of the number of shares sold at $t = 1$). Below (and in more detail in Section 4.3) we find that the HFT has both a negative and a positive effect on MM's revenue, which exactly offset each other. Thus, in this model HFTs have neither a direct nor an indirect effect on the amount of LT1's initial liquidity need i that is immediately executed–leaving "fill rates" (the percentage of the desired trade executed) and "trade delays" (the time to completion of desired trade) unaltered.

Liquidity can also be measured in terms of the cost of executing trades. We first look at average price paid and received; and then we look at the total revenue and total cost for each type of trader when selling and buying shares. As benchmark we use market clearing prices in the absence of an HFT. In that case the market clearing prices (which are the same for buyers and sellers) are:

$$P_1^{\Delta=0} = \mu - \frac{ia\sigma^2}{M+1}, \quad P_2^{\Delta=0} = \mu_2 \,.$$

This implies that without the HFT, LT2 faces infinite liquidity (buys at the asset's fundamental value), while LT1, the originator of the trade imbalance, faces a liquidity discount at date $t = 1$ of $\frac{ia\sigma^2}{M+1}$ (given to MMs).

When the HFT is present we observe two effects on average execution price: traders pay the HFT a haircut on a fraction of their trades, and the liquidity discount increases. We start with LT2's simpler problem, as he buys shares at their fundamental value. LT2 wants to buy i units (trade i) and ends up buying $\tilde{\theta}_2^{LT2}$ at $P_2 + \Delta_L$ per share, followed by $i + \theta_2^{LT2} - \tilde{\theta}_2^{LT2}$ at P_2 per share. Then, using the equilibrium values of P_2, θ_2^{LT2} and $\tilde{\theta}_2^{LT2}$, the average price paid by LT2 is:

$$\begin{aligned} \bar{P}_2^{LT2} &= P_2 - \frac{\tilde{\theta}_2^{LT2}\Delta_L}{i} = \mu_2 + \frac{(\Delta_L)^2}{ia\sigma^2} \,. \\ &= P_2^{\Delta=0} + \left(\frac{1}{4}\left(1 - \frac{1}{M+1}\right)\right)^2 ia\sigma^2 \,. \end{aligned} \tag{10}$$

From Equation (10), we can see that LT2 pays more to acquire his position than when there is no HFT. The price uplift payed by LT2 is given by the second term on the right-hand side of equation (10) which is the surplus lost to the HFT.

MMs trade twice. At date 1 they buy the asset, receive the liquidity discount and pay the haircut to the HFT; at date 2 they sell and pay the haircut. The average purchase price at date 1 is:

$$\begin{aligned} \bar{P}_1^M &= P_1 + \frac{\tilde{\theta}_1^M}{\theta_1^M}\Delta_S = \mu - \frac{ia\sigma^2}{M+1} - \frac{1}{4}\frac{ia\sigma^2}{M+1} \\ &= P_1^{\Delta=0} - \frac{1}{4}\frac{i}{M+1}a\sigma^2\,, \end{aligned} \tag{11}$$

from which we see that MMs receive an even better price than without HFT—they manage a 25% higher liquidity discount, even after accounting for the haircut paid to the HFT. Nevertheless, at date 2, the average sale price they obtain is:

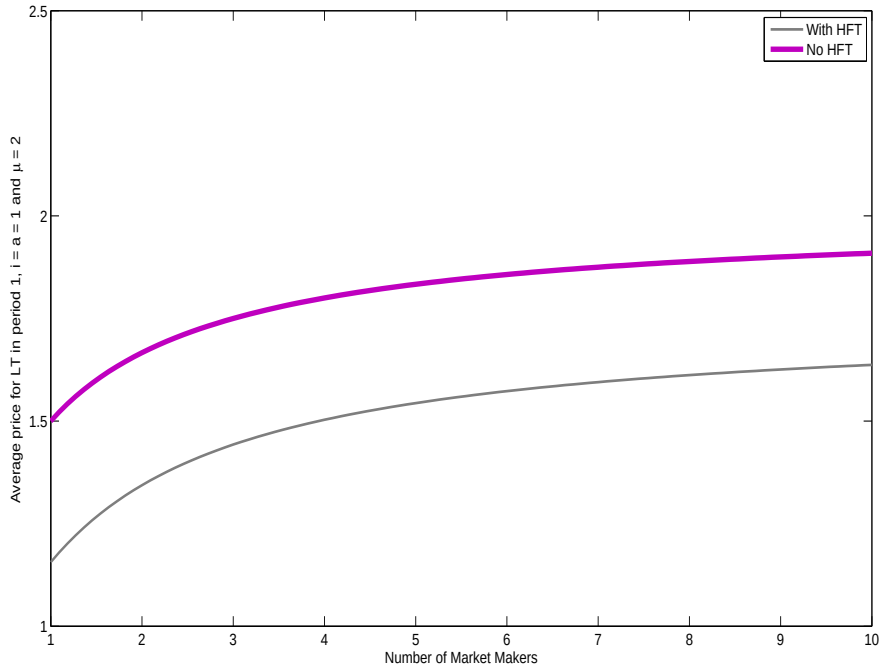
$$\bar{P}_2^M = \mu - \Delta_S + \frac{\tilde{\theta}_2^M}{\theta_1^M}\Delta_S = P_2^{\Delta=0} - \frac{1}{4}\frac{ia\sigma^2}{M+1}\,, \tag{12}$$

so they end up selling back at a lower price (relative to the case without HFT), and losing the initial extra liquidity discount garnered from the first transaction.

Finally, LT1 offers a liquidity discount and pays a haircut at date 1. In exchange he receives $\bar{P}_2^{LT1}$ when he completes his liquidity trade at $t = 2$. The average price received by LT1 at date 1 is:

$$\begin{aligned} \bar{P}_1^{LT1} &= P_1 - \frac{i - \tilde{\theta}_1^{LT1}}{i - \theta_1^{LT1}}\Delta_L \\ &= P_1^{\Delta=0} - \frac{1}{4}\left(\frac{M^2 + 5M + \frac{5}{4}}{M^2 + 2M + 1}\right) ia\sigma^2\,, \end{aligned} \tag{13}$$

Figure 4: Average price for LT1 in period 1



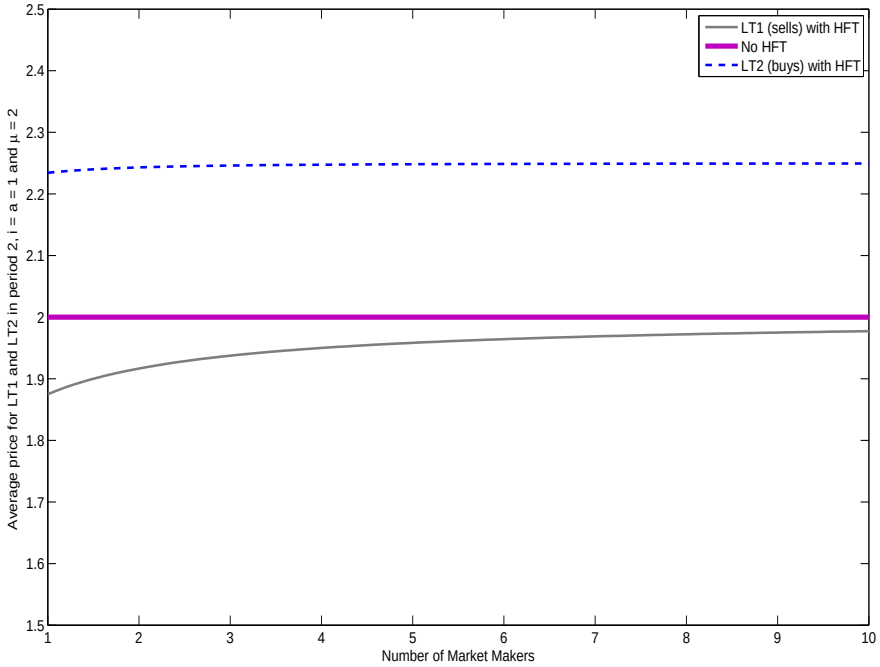
and at date 2:

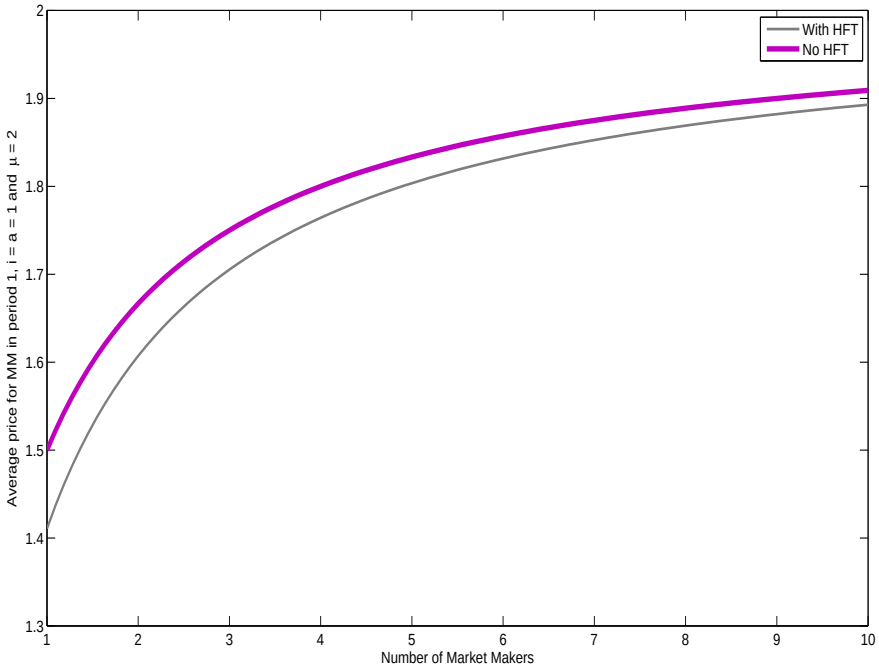
$$\begin{aligned} \bar{P}_2^{LT1} &= \frac{\mu_2\theta_1^{LT1} - \left(\theta_1^{LT1} - \tilde{\theta}_2^{LT1}\right)\Delta_S}{\theta_1^{T1}}, \\ &= P_2^{\Delta=0} - \frac{1}{4}\frac{ia\sigma^2}{M+1}. \end{aligned} \tag{14}$$

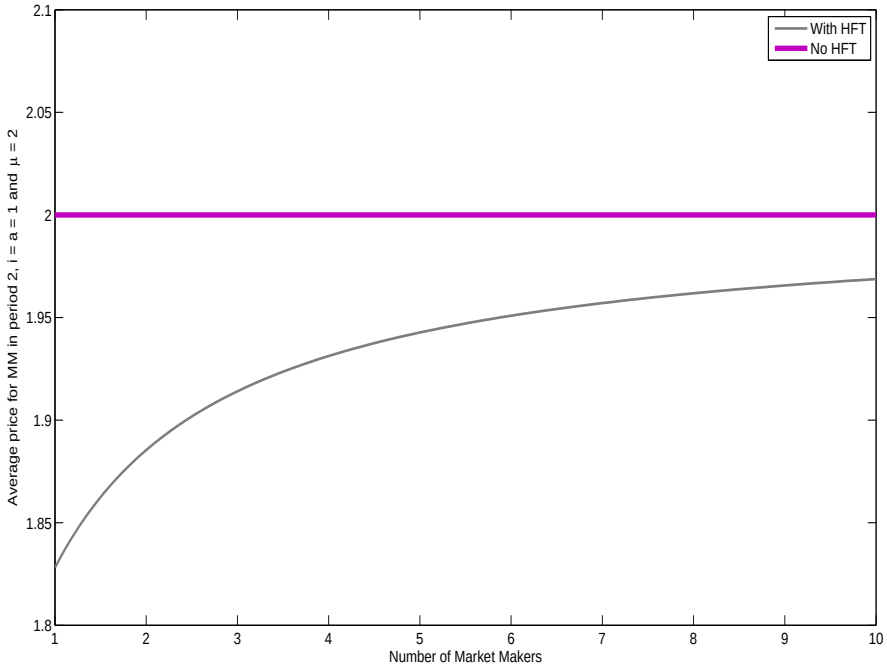
This implies that LT1 receives less money for liquidating his position at both dates (relative to the case without HFT) as can be seen in Figure 4.

From equations (10-14), it is clear that liquidity traders, LT1 and LT2, bear the brunt of the haircut imposed by the monopolist HFT, and that this impact is increasing in i. Although MMs suffer the loss of consumer (date 1) and supplier (date 2) surpluses, the liquidity discount received by the MMs for shares at date 1 is greater (the equilibrium price is lower) than in the absence of the HFT, and the net effect on his expected wealth is zero.

Figure 5: Average price for LT1 (sells) and LT2 (buys) in period 2







We now consider the effect on total revenues and costs. Clearly, if MMs face no average price impact, MMs' total revenue from intermediation is unaffected by the presence of HFTs. On the other hand, LT1's total revenues decrease in the presence of the HFT due to an increase in the liquidity discount and the loss of surplus. Similarly, LT2 faces a higher cost of acquiring his desired position as a result of the loss of surplus. But, if we compare which of the two liquidity traders is worse off we find that the effect of HFTs on LT1's revenue from selling i is the same as that on LT2's costs from buying i.

From Equations (13) and (14) we can obtain LT1's revenues with and without HFT:

$$R_{LT1} = \mu i - \frac{i^2 a\sigma^2}{(M+1)^2}\frac{1}{16}\left(4M^2+24M+3\right) \qquad \text{and} \qquad R_{LT1}^{\Delta=0} = \mu i - \frac{M}{(M+1)^2} i^2 a\sigma^2 \,.$$

Then, the presence of the HFT reduces his revenue from selling i shares by the amount

$$R_{LT1}^{\Delta=0} - R_{LT1} = \frac{i^2 a\sigma^2}{(M+1)^2}\frac{1}{16}\left(4M^2+8M+3\right) \,, \tag{15}$$

which increases, at an increasing rate, with the size of i and also increases in M.

Similar calculations give us LT2's costs with and without HFT:

$$C_{LT2} = \mu i + \frac{2M+1}{4(M+1)} i^2 a\sigma^2 - \frac{i^2 a\sigma^2}{(M+1)^2} \frac{1}{16} \left(4M^2 + 4M + 1\right) \qquad \text{and} \qquad C_{LT2}^{\Delta=0} = \mu i \,.$$

Interestingly, not only do LT2's costs of buying i shares increases in M, and in the size of the liquidity need, i (at an increasing rate), but the increase in cost from the HFT for LT2 is exactly the same as the reduction in revenue for LT1:

$$C_{LT2} - C_{LT2}^{\Delta=0} = \frac{i^2 a\sigma^2}{(M+1)^2} \frac{1}{16} \left(4M^2 + 8M + 3\right) = R_{LT1}^{\Delta=0} - R_{LT1}.$$

Finally, with a large number of MMs (as M tends to infinity), and using the fact that the liquidity discount (Equation (31)) becomes zero, we can characterize what happens to costs and revenues: $C_{LT2} \to \mu i + \frac{1}{4} i^2 a\sigma^2$ and $R_{LT1} \to \mu i - \frac{1}{4} i^2 a\sigma^2$ as $M \to \infty$, as well as what happens to HFT profits, which go to $\frac{1}{2} i^2 a\sigma^2$.

4.2 Price volatility

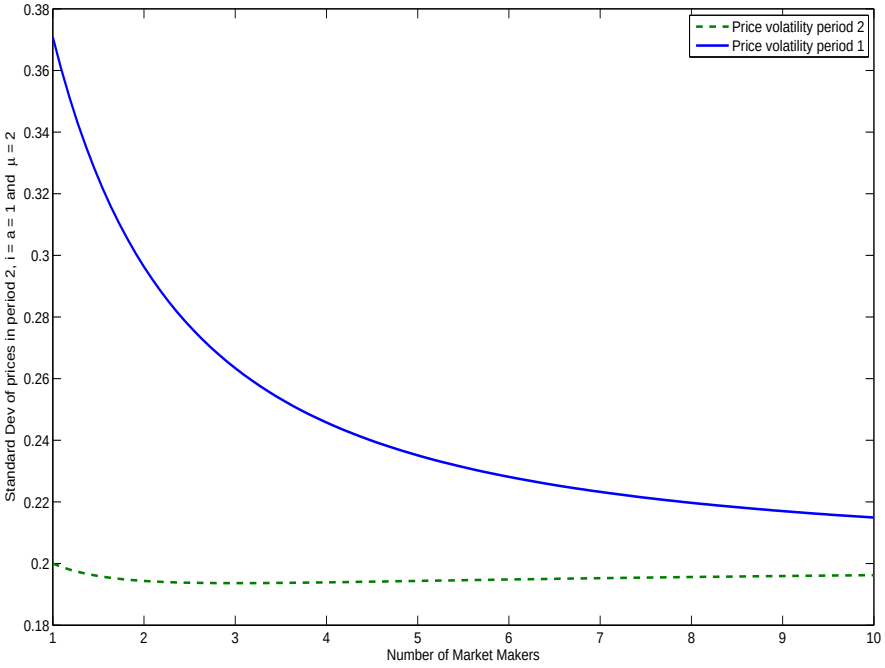
When there is no HFT intermediating transactions between LTs and MMs there is one transaction price at date 1, $P_1^{\Delta=0}$, and another transaction price at date 2, μ_2. However, when a monopolistic HFT intermediates all transactions the tape will record four prices in date 1, $\{P_1 - \Delta_L,\ P_1,\ P_1 + \Delta_S,\ P_1\}$, and four in date 2, $\{\mu_2 - \Delta_S,\ \mu_2,\ \mu_2 + \Delta_L,\ \mu_2\}$. Therefore, it is inevitable to observe price volatility within dates 1 and 2 which is solely caused by the HFT's presence. Figure 6 shows the volatility of prices in dates 1 and 2 that result from the HFT intermediating all trades.[15] Furthermore, Figure 7 depicts the standard deviation of prices at both dates. We see that price volatility is much higher when an HFT operates in the markets.

[15] To calculate the mean and variance of prices we also use the price at time $t = 0$ and assume it is the fundamental value μ, thus the mean and variance of the price in date 1 are

$$\bar{P}_1 = \frac{\mu + 4P_1}{5} \quad \text{and} \quad \mathbb{V}[P_1] = \frac{1}{5} \left\{ \left(\mu - \bar{P}_1\right)^2 + \left(P_1 - \Delta_L - \bar{P}_1\right)^2 + 2\left(P_1 - \bar{P}_1\right)^2 + \left(P_1 + \Delta_S - \bar{P}_1\right)^2 \right\} .$$

Similarly, to calculate the mean and variance of prices with date 2 we assume that the first observation is μ.

Figure 6: Volatility of prices in dates 1 and 2 induced by HF trading



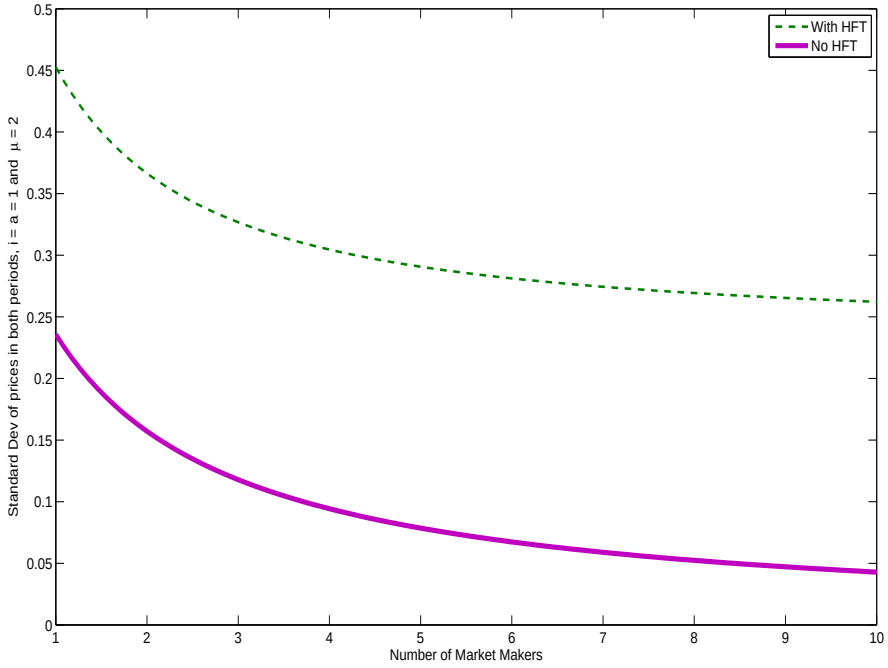
Note however that in our model this increased volatility does not generate additional risk for traders. Traders, in their evaluation of price risk (in their objective functions) recognize that the microstructure noise generated by the HFT translates into a deterministic effect on their execution costs while leaving their final holdings unaffected.[16]

4.3 Number of Market Makers

Our model allows us to derive the impact of HFT's surplus extraction activities on the number of MMs, and hence on the "true" supply of liquidity in the market. Although we know that HFTs have a zero net impact on the average price paid by MMs, and hence will not affect their revenues, we proceed as in GM and assume that the cost of entry for the MM is c. This allows us to analyze in detail the countervailing effects of the HFT on MMs. The entry cost

[16]This changes if we allow for the HFT to (randomly) miss the opportunity to intermediate some trades between LT and MMs. While adding such a feature will make the model more realistic, it complicates the analysis substantially without much additional insight.

Figure 7: Standard deviation of prices in dates 1 and 2 with and without HF trading



c is sunk before knowing the liquidity shock i which we assume to be normally distributed and independent of the other shocks affecting prices. At time $t = 0$ the expected utility of an individual MM is $\mathbb{E}\left[U\left(W_3^{MM} - c\right)\middle|\mathcal{F}_0\right]$, where W_3^{MM} is given by (23). Free entry of MMs will occur until

$$\mathbb{E}\left[U\left(W_3^{MM} - c\right)\middle|\mathcal{F}_0\right] = \mathbb{E}\left[U\left(W_0^{MM}\right)\middle|\mathcal{F}_0\right], \tag{16}$$

and recall that $\theta_2^{MM} = 0$. To calculate the expectations in equation (16) we require the expected value and variance of wealth in period three. The expected value of wealth can be decomposed into a liquidity discount and a loss of surplus:

$$\begin{aligned}
\mathbb{E}\left[W_3\middle|\mathcal{F}_1\right] &= W_0^{MM} - c \\
&\quad + \theta_1^{MM}\frac{ia\sigma^2}{M+1} + \theta_1^{MM}\Delta_S, \qquad \text{liquidity discount,} &(17)\\
&\quad - \theta_1^{MM} 2\Delta_S + \frac{1}{a\sigma^2}\left(2\left(\Delta_S\right)^2\right), \qquad \text{loss of surplus.} &(18)
\end{aligned}$$

From equation (17) we observe that the MM receives a liquidity discount for buying θ_1^{MM} shares from period 1 to 2. This discount consists of two terms. The first term $ia\sigma^2/(M+1)$ is the liquidity discount (per share) that an MM receives in the absence of an HFT; and the second term Δ_S is the extra discount per share that an MM receives if there is an HFT intermediating all trades. On the other hand, equation (18) shows the revenue loss from the HFT's surplus extraction.[17]

As for the variance of wealth, it remains unaffected by the presence of the (faultless) HFT:

$$\text{Var}\left[W_3^M \middle| \mathcal{F}_1\right] = \sigma^2 \left(\theta_1^{MM}\right)^2 . \tag{19}$$

The extra liquidity discount in the presence of an HFT (second term in equation (17)) exactly offsets the haircuts (18) paid to the HFT in the round-trip trade (buy from LT1 in period 1 and sell to LT2 in period 2), and the number of MMs will be the same with or without HFT.

The exact canceling of the two effects occurs because of the particular information assumptions we have made which allows the HFT to distinguish only large from small trades. In the Appendix we discuss the case where the HFT does not distinguish trades and charges the same haircut to all participants. In that case, the number of MMs increases, as the extra liquidity discount effect is greater than the haircuts.

5 Competition

We have studied the impact of a monopolistic HFT that extracts trading surplus in transactions between LTs and MMs. As the HFT profits substantially from this activity it should attract competition, specially from MMs who understand the usefulness of the additional speed and processing advantages. In fact there are several HFTs in the market and they represent a very small fraction of active traders (out of the almost 12,000 active traders during the Flash Crash,

[17]Note that there are two terms in equation (18). The first term is surplus extraction if the HFT took a haircut in periods 1 and 2 for all shares that the MM buys and sells, but recall from above that in every period the HFT breaks the MM's order in two batches and only applies a haircut to one batch. Therefore, the second term in equation (18) is a correction term that accounts for the quantities which are traded in both periods without paying a haircut.

Kirilenko et al. [2010] identify 16 as HFTs). Several reasons are put forth for the limited number of HFTs, the main being the high initial investment costs (costs of co-location and hardware, but more importantly, access to algorithms and detailed data) which act as effective barriers to entry.

The picture one obtains from the literature and discussions with market participants is that the number of HFTs has increased but that the potential profits for the latest and future entrants have dwindled to the point of being essentially negative. Despite this increase in the number of HFTs, profits from early entrants seem to remain stable. To capture this, we model competition between HFTs as a jockeying for position to become the monopolist for every (liquidity) trade. Thus, once an LT enters the market, HFTs compete amongst themselves to be the one that becomes the monopolist, and their success in this competition depends on their (relative) skills. Then, early entrants can continue to extract profits as long as they can retain their skill advantage at capturing rents from HFT trading.

We model competition in HF activities in two stages. First, we describe the outcome of competition between a fixed number of n HFTs who compete to extract surplus when a liquidity trade of size i is brought to the market. Second, we consider the entry decision of a potential new HFT—for example an MM who considers whether or not to become an HFT—who to enter must make an initial investment (sunk cost) to be able to compete with other HFTs.

Suppose there are n HFTs. Each trader has made an investment to acquire the resources and skills to participate in this market (which include hardware, software, human capital, co-location rights, etc.). Consequently, HFT j acquires skills described by a parameter η_j, which captures the effectiveness of the firm's investments as well as skills developed over time as a market participant. The higher the parameter η_j the higher is the effectiveness of HFT j to intermediate liquidity trades and extract surplus. Suppose that an LT enters the market with a liquidity need of size i. The potential surplus to be extracted by the HFT who becomes the monopolist is $\Pi(i)$ (given in Equation (8)). HFTs compete for this potential surplus, and their success in extracting it from this particular trade is determined by the realization (θ_j) of a random variable, Θ_j. To keep analytical tractability, we assume that Θ_j is exponentially distributed with parameter η_j, i.e. $\Theta_j \sim \exp(\eta_j)$, and the trader with the smallest realization

"wins the trade"–becomes the monopolist. Thus, for trader j, the probability of becoming the monopolist is

$$p(\eta_j) := \Pr\{\Theta_j = \min\{\Theta_1, \ldots, \Theta_n\}\} = \frac{\eta_j}{\sum_{u=1}^{n} \eta_u}.$$

For simplicity, we assume that every time there is a liquidity trade of size i, only one HFT "wins" and gets the whole surplus, the others get nothing.[18] Hence, the expected profit for trader j from a trade of size i is $p(\eta_j)\Pi(i)$. The reader can readily verify that if one were to introduce the parameter $p(\eta_j)$ into the optimal haircut problem for the monopolist in Section 3 none of the above results will be altered. We can now analyze the decision of a potential entrant.

Suppose an MM observes the activities of HFTs and considers the possibility of becoming an HFT himself. He studies the logistics of becoming an HFT and identifies the irreversible investment costs of setting up an HF trading desk (hardware, software, human capital, co-location, etc.). Let K be the total sunk costs of setting up and running the HF trading desk, including the opportunity cost of abandoning current activities (such as traditional market making). The potential entrant does his research and determines that if he were to obtain a skill level η from his investment, his expected profits would depend on $p(\eta)$, as well as the expected number of liquidity trades of size i that enter the market $N(i)$, and the profit each liquidity trade of size i, $\Pi(i)$. Let $\mathbb{E}[U(\Pi(i); N(i))]$ denote the MMs expected utility from future HFT profits. Then, an MM with skill η compares the utility from wealth K, $U(K)$, with the expected utility from future profits: $p(\eta)\mathbb{E}[U(\Pi(i); N(i))]$. Let $k = U(K)/\mathbb{E}[U(\Pi(i); N(i))]$ denote the relative "utility" cost of a trade. Then, the MM will become an HFT if

$$k \leq p(\eta).$$

The entrant has a reasonable estimate of k, but faces uncertainty with respect to the effectiveness of the investment, η. For simplicity we assume that the skill of the entrant, $\eta \in (1, \infty)$, is random and with a distribution such that its inverse, $1/\eta$, is uniformly distributed, that is $\frac{1}{\eta} \sim U(0, 1)$. As discussed above, the probability of becoming the HFT that intermediates a

[18] It is possible to incorporate additional complexity so that each HFT (and possibly the LT and MMs) gets a different fraction (possibly negative) of the surplus but it would not add any additional insights.

liquidity trade depends not only on the HFTs skill but more importantly, on his skill relative to those of existing HFT traders. Thus, any potential entrant needs to take into account that there already are n HFTs with skills described by parameters $\{\eta_u\}_{u=1}^n$. Then, an entrant with skill η will become the monopolist with probability

$$p(\eta) = \frac{\eta}{\eta + \sum_{i=1}^n \eta_i} = \left(1 + \frac{1}{\eta}\sum_{i=1}^n \eta_i\right)^{-1}.$$

Let $\alpha_j = \sum_{u=1}^n \eta_u > 1$ denote the total skill pool of other existing HFTs (excluding j)–we drop the j subscript until the final part of this section when it becomes useful. Then, the cdf and pdf of $p(\eta)$ are $G_p(z) = \frac{1}{\alpha}\left(1 + \alpha - \frac{1}{z}\right)$ and $g_p(z) = \frac{1}{\alpha}\frac{1}{z^2}$ respectively, and $\mathbb{E}[p(\eta)] = \frac{1}{\alpha}\log(1+\alpha)$.[19]

Given a relative cost of entry of k, and letting $\hat{\alpha}$ denote the total skills of incumbent HFTs, there will be entry until

$$k \geq \frac{1}{\hat{\alpha}}\log(1+\hat{\alpha}).$$

Clearly, incumbent HFTs suffer from entry, as for any HFT j greater competition increases the pool of skills of competing HFTs, α_j, which reduces her expected profits. Differences in profitability between early and later entrants can be due to skill differences that have not been eroded over time. As player's relative skills stabilize, the market settles around a core set of players who extract profits from a large fraction of trades, while less skilled, later entrants compete for the remaining trades and scrape enough to cover their operating (and opportunity) costs.

[19] See the Appendix for details.

6 Conclusions

We use the GM model as becnhmark to introduce an HFT who due to her rapid execution and information processing ability is able to intermediate trades between liquidity traders and market makers. In our model the HFTs devise strategies which are carefuly tailored to extract trading surplus from market participants and to hold zero inventories across periods.

Our model shows that the presence of HFTs increases the price impact of liquidity trades (in proportion to the size of the trade), has no effect on the number of market makers (although this effect depends on the way HFTs choose to extract surplus from the other traders), increases price volatility, and doubles trading volume. Thus, the presence of HFTs distorts market conditions, not through the amount of shares traded, but through prices. By exacerbating price impact, the HFT induces additional market impact costs on participants, especially the liquidity trader. This cost is proportional to the size of the trade which implies that large liquidity traders, including institutional investors trading to change the composition of a portfolio, are the most affected by the presence of HFTs–an effect that is consistent with Zhang [2010]'s findings.

In the particular case we have analyzed, it is the liquidity traders who bear all the costs from the presence of the HFT. Liquidity traders lose on two accounts: they lose trading surplus to the HFT, and they must offer a higher liquidity discount to market makers to get them to buy their shares. On the other hand, market makers find themselves unaffected because the revenue they lose to the HFT is compensated by a higher liquidity discount from liquidity traders. This implies that overall, the HFTs reduce the value of the stock market as a forum for providing a way for investors to convert their equity into cash (and viceversa) quickly and at a reasonable price falls because of the adverse effect the HFT has on prices. This value reduction would, in a more general framework, be passed on to the firms raising capital in equity markets. As equity buyers recognize the increased trading execution costs from HFTs, they will require greater discounts from IPO issuers, resulting in greater IPO underpricing, specially for large institutional investors.

An aspect we have left unmodeled is the time to execution faced by liquidity traders. Clearly, if the HFT is to intermediate between liquidity traders and market makers, it must

act quicker and hence execution time for liquidity traders must be lower. SEC [2010] reports a reduction in average execution time from 10.1 seconds in January 2005 to 0.7 seconds in October 2009. Whether the increased execution speed compensates for the additional trading costs is something that requires a more detailed analysis, but also it is something that traders are facing little choice on.

Another issue that arises from our analysis is the question of how to measure (socially valuable) liquidity, when the analyst has no access to the identity of traders and hence cannot determine directly the market impact costs of trades. Clearly, just adding up the number of trades in the presence of rent-seeking hyperfast algorithms whose asset positions are essentially zero most of the time seems like a poor measure of the ability of the market to provide prompt and fair value to investors. Also, HFTs who generate additional microstructure noise at smaller intervals and accelerate market transactions raises several questions regarding how to measure price impact, whether we should adjust current measures to account for the increase in the speed of execution, and whether these measures adequately capture an investor's cost of executing a trade.

Finally, although in our knife-edge case HFTs had no effect on market makers, in the Appendix we consider an alternative strategy where HFTs always charges the same haircut per trade (regardless of the size of the trade, type of trader, and date), and the effect is to increase the number of market makers. This suggests that the overall effect on the "true liquidity providers" in not at all clear, and we need good empirical work to (a) determine the speed-cost effect on outside investors and liquidity traders, (b) determine if HFTs are raising the cost of business for market makers, and (c) if they do, whether the liquidity they provide is a good substitute for the one that is being driven out. HFTs clearly generate costs, but they also generate benefits, and the net effect requires more empirical analysis.

References

Jonathan Brogaard. High frequency trading and its impact on market quality. *SSRN Working Paper*, 2010.

Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. Recent trends in trading activity. *SSRN Working Paper*, 2010.

U.S. Commodity Futures Trading Commission, the U.S. Securities, and Exchange Commission. Findings regarding the market events of may 6, 2010. Report, SEC, September 2010.

Jaksa Cvitanić and Andrei A. Kirilenko. High Frequency Traders and Asset Prices. *SSRN eLibrary*, 2010.

David Easley, Marcos Mailoc Lopez de Prado, and Maureen O'Hara. The microstructure of the 'Flash Crash': Flow toxicity, liquidity crashes and the probability of informed trading. *The Journal of Portfolio Management*, 37(2):118–128, November 2011.

Sanford J. Grossman and Merton H. Miller. Liquidity and market structure. *Journal of Finance*, 43(3):617–37, July 1988.

Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld. Does algorithmic trading improve liquidity? *Journal of Finance*, Forthcoming, 2010.

Michael Kearns, Alex Kulesza, and Yuriy Nevmyvaka. Empirical limitations on high frequency trading profitability. *SSRN worjing papers*, 2010.

Andrei A. Kirilenko, Albert (Pete) S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The Flash Crash: The Impact of High Frequency Trading on an Electronic Market. *SSRN eLibrary*, 2010.

SEC. Concept release on equity market structure. Concept Release No. 34-61358; File No. S7-02-10, SEC, January 2010. 17 CFR PART 242.

Frank Zhang. The Effect of High-Frequency Trading on Stock Volatility and Price Discovery. *SSRN eLibrary*, 2010.

A Appendix

Above we argued that the HFT can discriminate across order size when trades come to the market which enables her to apply haircuts for large and small trades. If the HFT is not able to discriminate by size or type of trader she can still exercise her monopoly power by applying one haircut to all trades she intermediates. Hence we can repeat the analysis above while setting the following haircuts:

$$\Delta = \Delta_1^{T1} = \Delta_2^{T1} = \Delta_1^{MM} = \Delta_2^{MM} = \Delta_2^{T2}\,.$$

Then, the optimal delta set my the HFT is:

$$\Delta = ia\sigma^2 \frac{(2M+1)}{(M+1)(3+2M)}\,. \tag{20}$$

As in the case with two haircuts discussed in the body of the paper, LT1's optimal holding is $\theta_1^{T1} = i/(M+1)$, but as we show below, the equilibrium number of MMs in this case is *higher* than without HFT.

Other results are similar to those discussed above: First, the volatility of realized transaction prices increases. Second, the price impact of the liquidity trades in both periods is substantial: equilibrium sell (buy) prices are lower (higher) than the competitive price in the absence of the HFT. Third, the expected returns that the MM face from buying in period 1 and selling in period 2 increase (at the expense of traders), and the equilibrium number of MMs present in the market increases when compared to the number of MMs that are present in a market without a HFT. Fourth, the total volume of trades doubles relative to the number of trades observed in the absence of the HFT.

In the interest of space we only discuss the equilibrium number of MMs. We proceed as above where the entry condition (16) becomes

$$e^{ac}\mathbb{E}\left[e^{-\frac{1}{2}a^2\sigma^2\left(\frac{i}{M+1}\right)^2\beta(M)}\right] \;=\; 1 \qquad \text{where} \qquad \beta(M) = \frac{12M^2+12M+7}{(3+2M)^2}\,. \tag{21}$$

By inspecting (21) we know that the number of MMs with and without an HFT will be the same only when $\beta(M) = 1$ and this occurs for $M = 1/2$. When $M > 1/2$ we have more MMs in the presence of HFT. If we denote the number of MMs by M and the number of MMs in the absence of HFT by $M_{\Delta=0}$ we can show that

$$M_{\Delta} = \sqrt{\beta\left(M\right)}\left(M_{\Delta=0}+1\right)-1\,,$$

hence we have that $M > M_{\Delta=0}$. (The function $\beta(M)$ is increasing in M and we are interested in values for $M > 1$).

A.1 General Haircuts

$$\begin{aligned}\Pi_2\left(i\right) &= \left|\tilde{\theta}_2^{LT2}\left(P_2+\Delta_2^{LT2}\right)+i\right|\Delta_2^{LT2}+\left|\tilde{\theta}_2^{LT1}\left(P_2-\Delta_2^{LT1}\right)-\theta_1^{LT1}\right|\Delta_2^{LT1}\\ &\quad +M\left|\tilde{\theta}_2^{MM}\left(P_2-\Delta_2^{MM}\right)-\theta_1^{MM}\right|\Delta_2^{MM}. \qquad (22)\end{aligned}$$

$$\begin{aligned}W_3^{LT1} &= W_0^{LT1}+\theta_2^{LT1}P_3+\left(\theta_1^{LT1}-\theta_2^{LT1}\right)P_2-\left(\theta_1^{LT1}-\tilde{\theta}_2^{LT1}\right)\Delta_2^{LT1}-\theta_1^{LT1}P_1+\left(i-\tilde{\theta}_1^{LT1}\right)\Delta_1^{LT1}\,,\\ W_3^{MM} &= W_0^{MM}+\theta_2^{MM}P_3+\left(\theta_1^{MM}-\theta_2^{MM}\right)P_2-\left(\theta_1^{MM}-\tilde{\theta}_2^{MM}\right)\Delta_2^{MM}-\theta_1^{MM}P_1-\tilde{\theta}_1^{MM}\Delta_1^{MM} \qquad (23)\end{aligned}$$

$$\theta_1^{LT1} = \frac{1}{a\sigma^2}\left(\mu-P_1-\Delta_2^{LT1}\right)\,, \qquad (24)$$

$$\theta_1^{MM} = \frac{1}{a\sigma^2}\left(\mu-P_1-\Delta_2^{MM}\right)\,. \qquad (25)$$

$$P_1 = \mu-\frac{\Delta_2^{LT1}+M\Delta_2^{MM}+ia\sigma^2}{M+1}\,, \qquad (26)$$

$$\theta_1^{LT1} = \frac{i}{M+1} + \frac{1}{a\sigma^2}\frac{M}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right), \tag{27}$$

$$\theta_1^{MM} = \frac{i}{M+1} - \frac{1}{a\sigma^2}\frac{1}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right), \tag{28}$$

$$\begin{aligned}\tilde{\theta}_1^{LT1} &= \frac{i}{M+1} + \frac{1}{a\sigma^2}\frac{M}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right) + \frac{\Delta_1^{LT1}}{a\sigma^2}, \\ \tilde{\theta}_1^{MM} &= \frac{i}{M+1} - \frac{1}{a\sigma^2}\frac{1}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right) - \frac{\Delta_1^{MM}}{a\sigma^2}.\end{aligned}$$

$$\Pi_1(i) = \left|i - \tilde{\theta}_1^{LT1}\left(P_1 - \Delta_1^{LT1}\right)\right|\Delta_1^{LT1} + M\left|\tilde{\theta}_1^{MM}\left(P_1 + \Delta_1^{MM}\right)\right|\Delta_1^{MM}. \tag{29}$$

For a given order imbalance of magnitude $i > 0$:

1. Market clearing prices are

$$\begin{aligned}P_1 &= \mathbb{E}[P_2|\mathcal{F}_1] - \frac{ia\sigma^2}{M+1} - \frac{\Delta_2^{LT1} + M\Delta_2^{MM}}{M+1}, \\ P_2 &= \mathbb{E}[P_3|\mathcal{F}_2] = \mu_2,\end{aligned} \tag{30}$$

and the magnitude of the liquidity discount at $t = 1$ is

$$\mathbb{E}[P_2|\mathcal{F}_1] - P_1 = \frac{ia\sigma^2}{M+1} + \frac{\Delta_2^{LT1} + M\Delta_2^{MM}}{M+1}. \tag{31}$$

   (a) Asset trading is described on Table 2:

Table 2: Prices and Volume of Trades

Price	LT1	LT2	MM (total)
$P_1 - \Delta_1^{LT1}$	$-\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + \frac{\Delta_1^{LT1}}{a\sigma^2}$		
P_1	$-\frac{\Delta_1^{LT1}}{a\sigma^2}$		$M\frac{\Delta_1^{MM}}{a\sigma^2}$
$P_1 + \Delta_1^{MM}$			$\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) - M\frac{\Delta_1^{MM}}{a\sigma^2}$
$P_2 + \Delta_2^{LT2}$		$i - \frac{\Delta_2^{LT2}}{a\sigma^2}$	
P_2	$-\frac{\Delta_2^{LT1}}{a\sigma^2}$	$\frac{\Delta_2^{LT2}}{a\sigma^2}$	$-M\frac{\Delta_2^{MM}}{a\sigma^2}$
$P_2 - \Delta_2^{LT1}$	$-\frac{1}{M+1}\left(i + M\frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + \frac{\Delta_2^{LT1}}{a\sigma^2}$		
$P_2 - \Delta_2^{MM}$			$-\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + M\frac{\Delta_2^{MM}}{a\sigma^2}$

(b) The HFT acts as counterparty to all these trades and makes profits equal to

$$\Pi(i) = \Pi_1(i) + \Pi_2(i) , \tag{32}$$

where $\Pi_1(i)$ and $\Pi_2(i)$ are described in Equations (29) and (22) respectively.

A.2 Competition

Result: The distribution of $p(\eta)$ has cdf $G_p(z) = \frac{1}{\alpha}\left(1 + \alpha - \frac{1}{z}\right)$ and pdf $g_p(z) = \frac{1}{\alpha}\frac{1}{z^2}$. Furthermore, $\mathbb{E}[p(\eta)] = \frac{1}{\alpha}\log(1+\alpha)$.

Proof: As $1/\eta \sim U[0,1]$ then $1/p(\eta) \sim U[1, 1+\alpha]$, and the support of the distribution of $p(\eta)$ is $\left[\frac{1}{1+\alpha}, 1\right]$. To compute the cdf of $p(\eta)$ we use the result that for a random variable X

with cdf $F_X(z)$ the cdf of the random variable $Y = g(X)$, where g is a deterministic decreasing function, is given by

$$Y \sim G(z) \text{ where } G(z) = 1 - F_X\left(g^{-1}(z)\right).$$

Thus, letting $g(x) = 1/x$ and using $F_X(k) = (k-1)/\alpha$, the cdf and pdf of $p(\eta)$, denoted by G_p and g_p respectively, are given by

$$\begin{aligned} G_p(z) &= 1 - F_X\left(\frac{1}{z}\right) \\ &= 1 - \frac{z^{-1} - 1}{\alpha} = \frac{1}{\alpha}\left(1 + \alpha - \frac{1}{z}\right), \end{aligned} \tag{33}$$

and by differentiating (33) with respect to z we obtain

$$g_p(z) = \frac{1}{\alpha}\frac{1}{z^2}.$$

Finally, it is straightforward to calculate

$$\mathbb{E}\left[p(\eta)\right] = \frac{1}{\alpha}\log(1+\alpha).$$

□